Quality preserving and everlasting databases – a retrospective

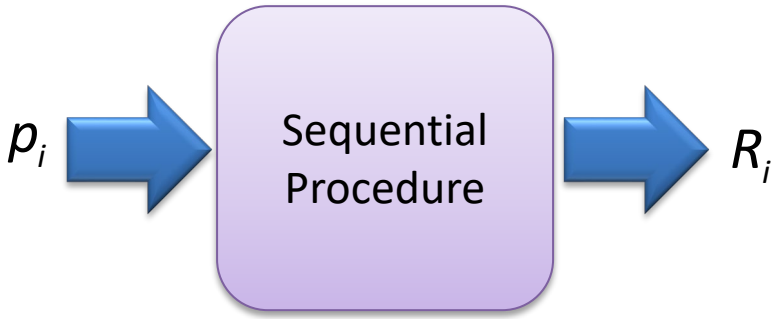
Saharon Rosset, Tel Aviv University

Loosely based on:
Generalized α-investing: definitions, optimality results and application to public databases
by Aharoni & R., JRSSB (2014)

Mapping the challenges of scientific discovery on shared data resources ("public databases")

- Basic challenge: preserve statistical validity of findings
  - Control some measure of overall false discovery
- Sequential use, future use not known in advance
  - Need sequential methods for false discovery control
- We want to make sure it will remain useful
  - Can keep using it in the future
  - Can keep making scientifically valid discoveries
- Different modes of use for scientific discovery
  - (Non-adaptive) Dependence between tests and queries
  - Adaptive use: future questions depend on previous answers

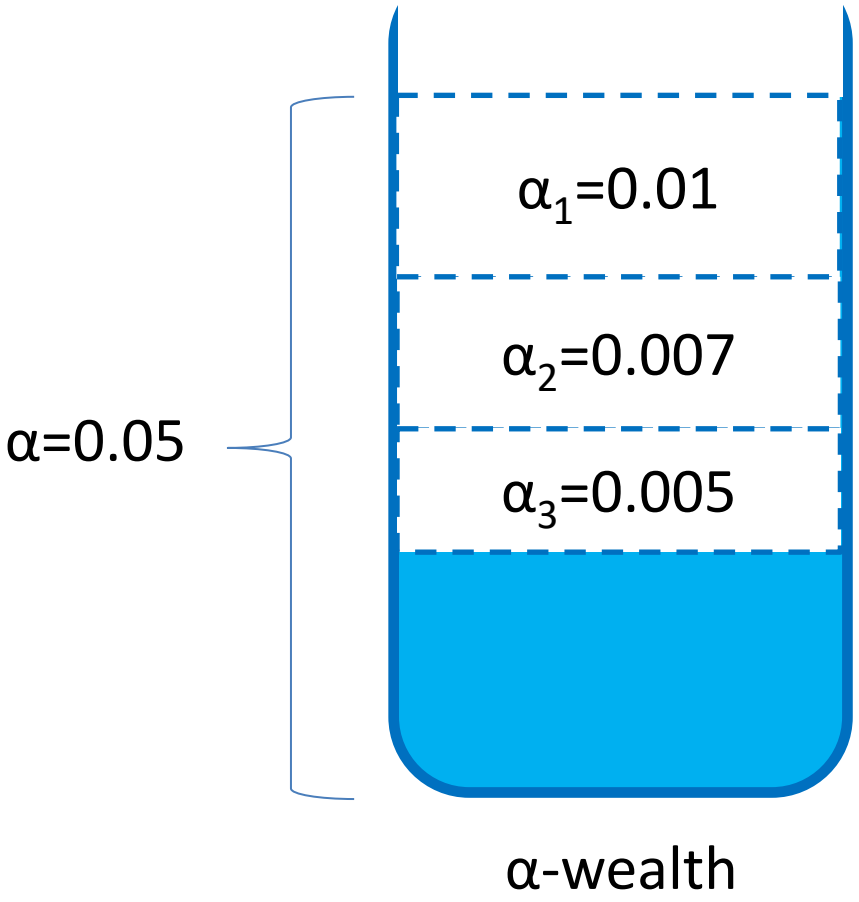
Sequential procedures



Guarantees:

(some overall measure of false discoveries) $\leq \alpha$

Example: Alpha Spending controls FWER



Test 1: $R_1=1$ if $p_1<\alpha_1$
Test 2: $R_2=1$ if $p_2<\alpha_2$
Test 3: $R_3=1$ if $p_3<\alpha_3$

.
.
.

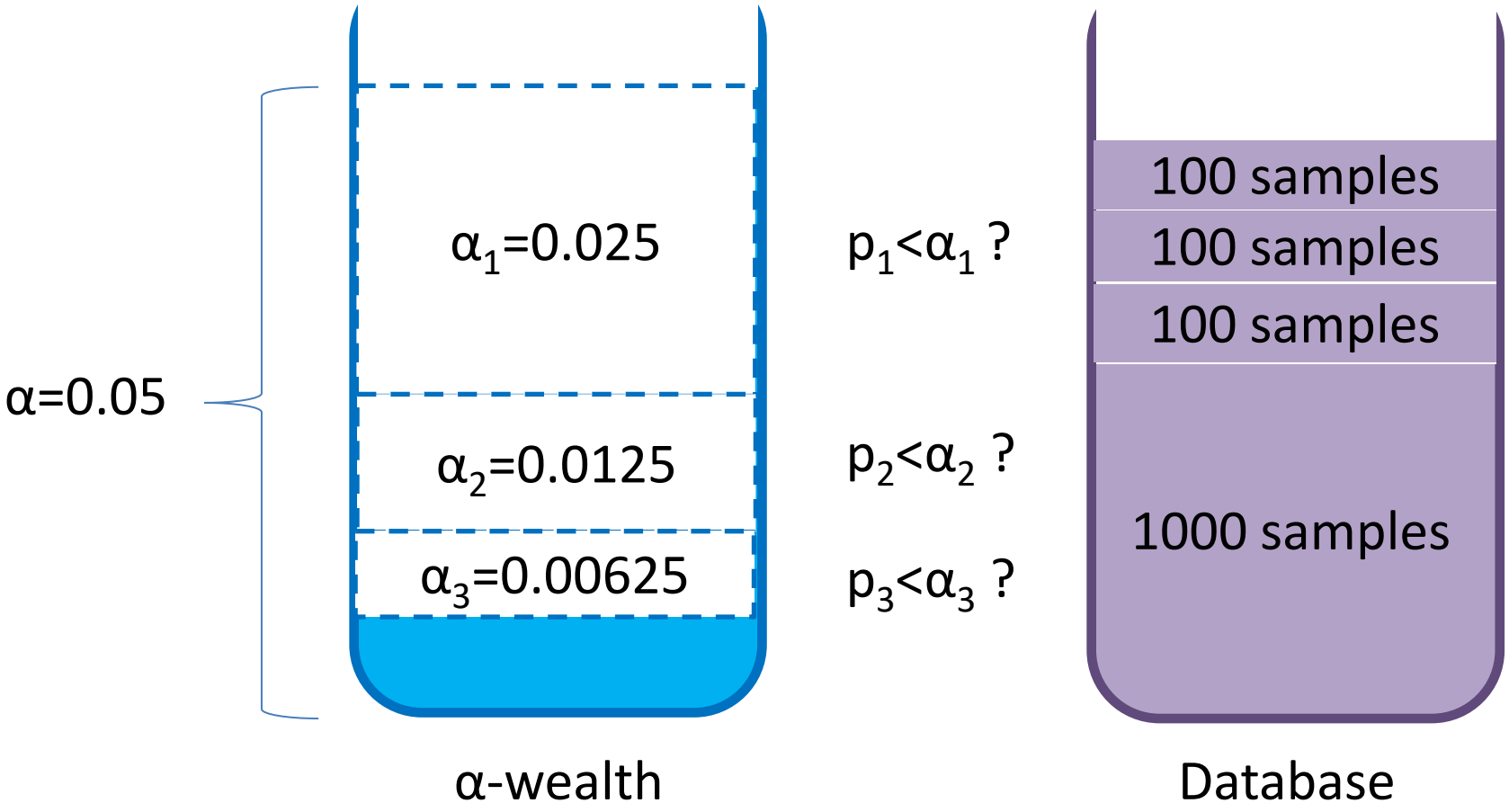
Condition: $\sum_{i=1}^{\infty} \alpha_i \leq \alpha$

The Quality Preserving Database (QPD)

- Motivation: as levels decrease, power also decreases
  - Earlier users are “using up” the data
- How can we compensate for usage?
  - “Stable” method: user does not lose power by arriving later
  - Two ways to achieve this:
    - Not decrease level (impossible)
    - Add samples $\Rightarrow$ more power at same level
- QPD basic problem: design “payment” schemes for usage which guarantee power to next users, while keeping costs bounded
  - Turns out to be possible in many cases

Aharoni, Neuvirth, R. (2012) TCBB
R., Aharoni, Neuvirth (2014) Gen. Epi.

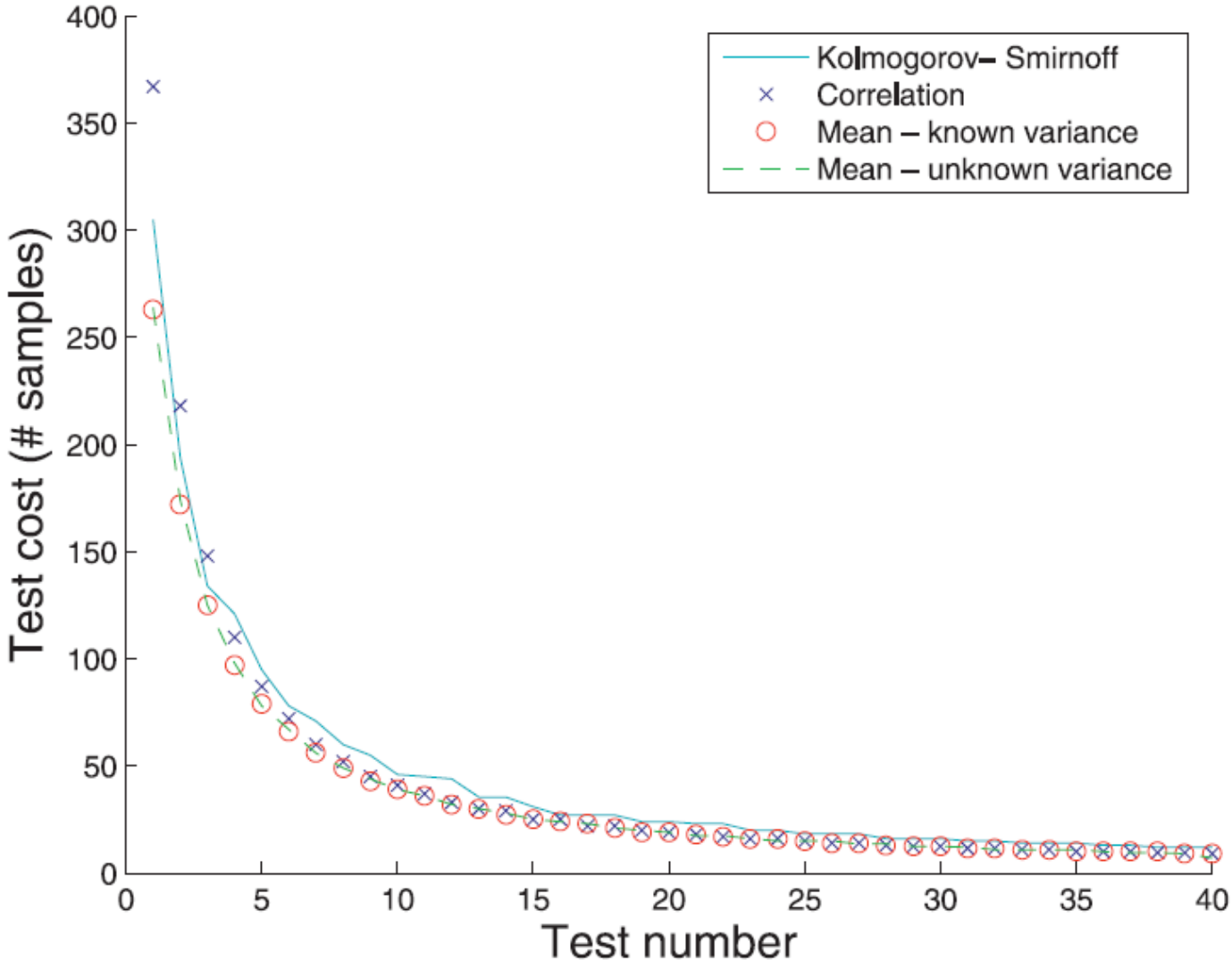
QPD schematic view



QPD implementation example

- Stream of normal tests, each with effect size s_i and power requirement π_i
- When test i arrives, we have n samples in the database
  - Find c such that $\alpha q^n(1-q^c)$ is sufficient level for obtaining power π_i
  - Request c samples (or equivalent cost) in payment from i
- This simple recipe guarantees “stability”
  - The ability to serve an infinite stream with bounded costs
- Applicable well beyond normal distributions
- In practice, it leads to quickly diminishing costs

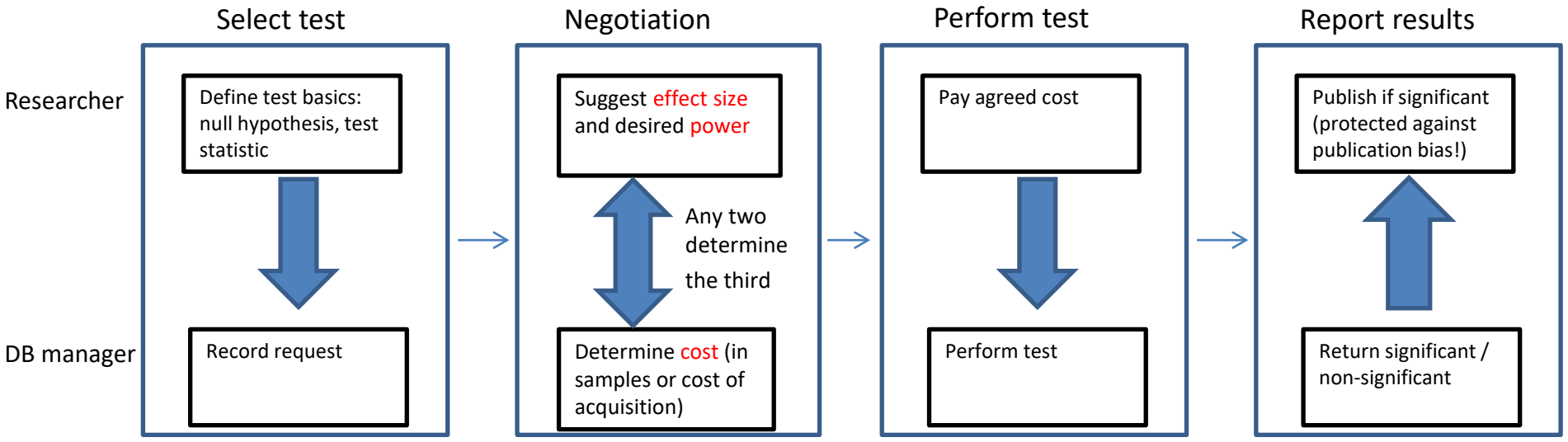
QPD simulations



Basic QPD Summary

- Tool for false discovery control in public databases, with a management layer responsible for validity
- Pay per use with samples or cost of sample acquisition
- Can serve an infinite series of requests without loss of power
- Uses Alpha Spending to control FWER
- Scientists concentrated on discoveries (effect size and power)
- Scientists do not see p-values, only R's (reject/not reject H0)

QPD schema


Select test
Negotiation
Perform test
Report results
Researcher
DB manager
Define test basics: null hypothesis, test statistic
Record request
Suggest effect size and desired power
Any two determine the third
Determine cost (in samples or cost of acquisition)
Pay agreed cost
Perform test
Publish if significant (protected against publication bias!)
Return significant / non-significant

QPD use case: GWAS replication server

Assume a community (e.g. Type 2 diabetes researchers) builds a QPD for replicating findings

- Initialize with, say, 500 samples

Comparison of different scenarios:

- Replicate on own data:
  Requires hundreds of samples, publication bias

- Replicate on public data:
  Requires no samples, but severe publication bias

- Use QPD:
  Requires <5 samples, protected from (replication) publication bias

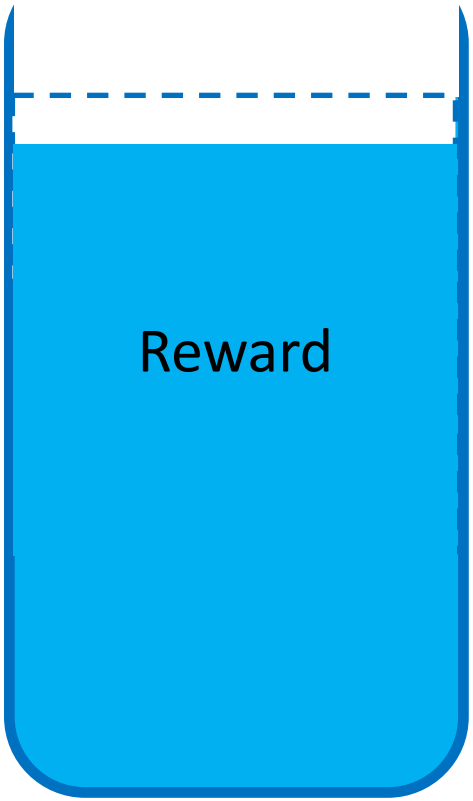
Sequential control beyond FWER: Alpha Investing (Foster & Stine 2007)

- Similar to Alpha Spending
- Guarantees mFDR ≤ α
- Requires “almost independence” of hypotheses

Marginal False Discovery Rate (mFDR): $\frac{E(\mathrm{V})}{E(\mathrm{R})+1-\alpha}$

V — the number of false discoveries
R — the total number of discoveries

Alpha Investing



✘ $p_1 < x_1/(1+x_1)$? Reward $= x_1/(1+x_1)+\alpha$

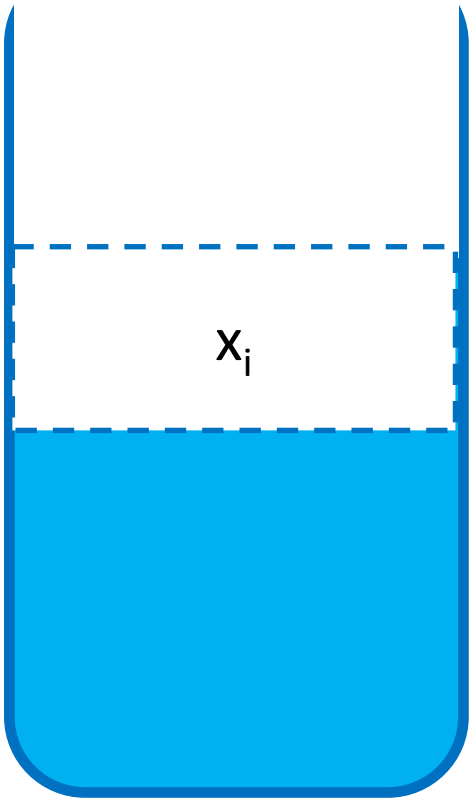
✘ $p_2 < x_2/(1+x_2)$? Reward $= x_2/(1+x_2)+\alpha$

✔ $p_3 < x_3/(1+x_3)$? Reward $= x_3/(1+x_3)+\alpha$

Summary of results from Aharoni & R. (JRSSB, 2014)

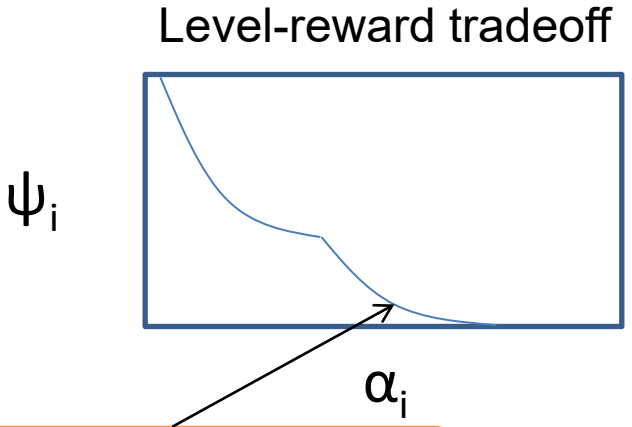
- Define Generalized Alpha Investing (GAI): what combinations of (level, reward) are legal to control mFDR
- Can find Optimal GAI: maximizing expected reward and hence future levels and power
- Can design variants that can be used within QPD (still under near-independence assumption)

Alpha Spending with Rewards



wealth

$\alpha_i = x_i, \quad \psi_i = \alpha$



Alpha Spending with Rewards

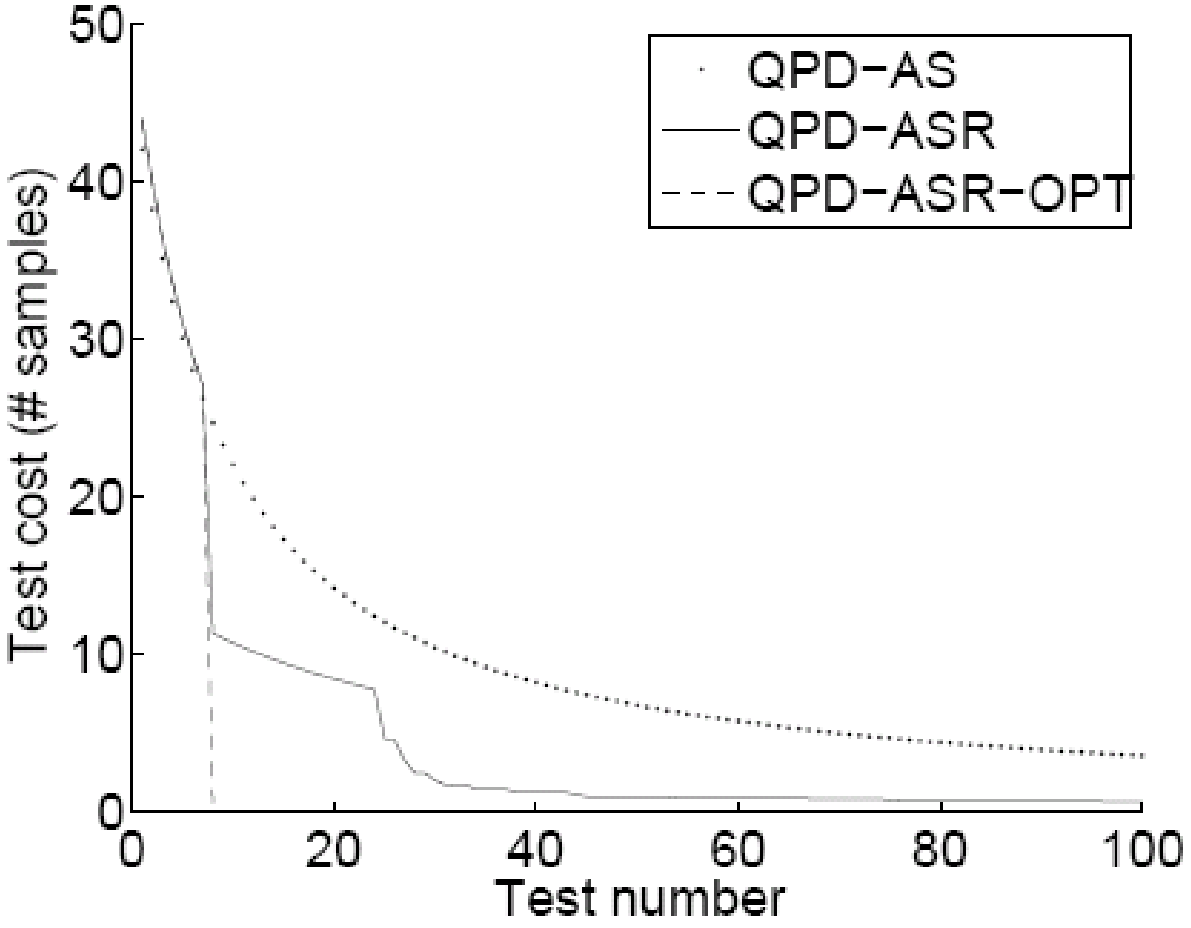
Generalized alpha investing and QPD

- Can we use these powerful sequential testing approaches to decrease QPD costs?
- As it turns out, Alpha Spending with Rewards is the one that can be integrated with QPD
- The rewards mean that costs decrease more quickly
  - In typical cases most costs are zero!

Simulation

- We simulated 100 requests for t-tests, where
  - Power=0.95
  - Effect-size=0.1
  - Probability of a true null 0.9
- Initial number of samples n_0=2000
- Three variants of QPD, all with:
  - α=0.05
  - q=0.999

Alpha Spending with Rewards and QPD



QPD and FDR control

- Javanmard and Montanari (2017) were the first to devise methods for sequential FDR control
- Their LOND method is a Generalized Alpha Investing rule (though not exactly Alpha Spending with Rewards)
  - Controls both FDR and mFDR
- It can be implemented within QPD
  - More conservative than QPD-ASR, but still gains compared to using Alpha Spending
- Lots of developments for sequential FDR control since ($\Rightarrow$ Aditya)

Dependence vs Adaptation

- Adaptive data analysis: Decide on next analysis given results of previous analyses
- In the context of hypothesis testing: decide on the next test based on the results of previous tests
- So it's not only the outcomes of tests that are dependent, but the selection of the tests performed is dependent on previous results
- QPD methods presented so far do not deal with adaptive tests

The reusable holdout

- Science paper by Dwork et al. (2015) shows how to reliably estimate $O(n^2)$ means with n data even if functionals are chosen adaptively
  - Motivation: evaluating Machine Learning models
  - However can be used as-is to guarantee similar results for hypothesis testing using Alpha Spending (⇒ FWER control)

- Problems:
  - The O() notation hides practical issues
  - What happens after it runs out?
  - Can "cost" of estimation be reduced from $1/\sqrt{n}$ if users are not adaptive?

The everlasting database

- Goals:
  - Maintain infinite usefulness
  - Allow “cheap” non-adaptive queries/tests
  - Adapt automatically to users’ adaptiveness
- Implementation, using and enhancing “reusable holdout” techniques:
  - Charge O(1/n) for query n
  - Return “slightly” noisy answers (=decrease levels of tests)
  - Monitor “overfitting”
  - When overfitting too much – renew data and charge O(n)

Woodworth, Feldman, R., Srebro, NIPS 2018

The everlasting database

- Guarantees (with high probability):
  - Non adaptive users pay O(log m) for m queries
  - Adaptive users pay $O(m^{1/2})$ on average because they will eventually trigger the penalty mechanism
  - Power is preserved for all users

- Practical issues:
  - Large constants in O()
  - Large initial data size needed
  - Unbounded costs (penalty mechanism)

Woodworth, Feldman, R., Srebro, NIPS 2018

Summary

- QPD: a paradigm for public database management
  - False discovery and publication bias control
  - No loss of power
  - Cost: (slow) augmentation of database size
- QPD can be implemented in practice:
  - Fits well with trends of both sharing and security/privacy
  - Requires change in how testing is done: no more p values!
  - Practical issues: cultural acceptance, data quality, gaming
- Challenges for both research and implementation:
  - Controlling different criteria while allowing dependence
  - Dealing with adaptive data analysis – can it be made practical?

Thanks!

saharon@tauex.tau.ac.il